Posudek oponenta diplomové práce

Jméno: Bc. Martin Majliš
Název: Velký mnohojazyčný korpus/Large Multilingual Corpus
Vedoucí: doc. Ing. Zdeněk Žabokrtský, Ph. D.

Práce Large Multilingual Corpus se zabývá vytvořením textového korpusu pro velké množství jazyků z dat, která se nacházejí na Internetu. Cílem práce bylo získat pro každý z jazyků alespoň 10 miliónů slov.

Výsledkem práce je jednak jazykový korpus (který ovšem není dostupný mimo pracoviště), jednak nástroje, které by měly umožnit takový korpus vytvořit.

Text je rozvržen do několika kapitol: přehled podobných projektů a metod používaných pro jednotlivé podúkoly, popis použitých metod a dosažených výsledků. Softwarová část je zpracována formou shell (bash) skriptů a programů v jazyku Perl.

Obsah práce odpovídá zadání, autor navrhl a implementoval sadu nástrojů, které slouží k vybudování korpusu, jedná se zejména o samotné stahování souborů formátu HTML, odstranění HTML značek a detekci jazyka.

K textu a obsahu práce mám následující připomínky:

- Detekce duplicit funguje velmi jednoduše tak, že zahazuje často se vyskytující řádky textu. Dále provádí náhodnou permutaci řádků textu, čímž dochází k roztržení vět, což může velmi omezovat následné zpracování a použití korpusů.
- Ukázkové korpusy mají problém se slitím po sobě následujících slov.
- Z textu práce není jasné, jak jsou řešeny problémy s různým kódováním (např. UTF-8 vs CP-1250).
- Architektura systému je zřejmě do značné míry ovlivněná aktuálními možnostmi studenta na MFF (vybavení laboratoří, ufallab atd.), je otázka, zda je systém dostatečně obecný.
- U jednotlivých kroků chybí popis HW nároků (zejména paměťových) a časové náročnosti. Není pak možné dobře posoudit, zda byla správně zvolena distribuovaná architektura systému.
- V textu se vyskytují překlepy drobné: (“hight quality” (str. 73); “Building web corpus was divided smaller jobs, …” (str. 29); “I wouldhave” (str. 46) a mnoho dalších), i faktické: (170 GB vs 170 TB (str. 27)).
- Chyby ve stylu a formátování (str. 6, 8, nejednotné označení: “UTF8”, “non-UTF8”, “UTF-8”, “utf-8”).
- Kapitola 4.5 Internet size: není jasné, jak autor dospěl k prezentovaným číslům.
- Popis tabulek také není vždy srozumitelný (Table 4.8 Internet – page counts, z popisku není jasné, v jakých je to jednotkách).
- Chybí popis adresáře SolA, popis zapojení boilerplate programu, …

Naopak oceňuji skripty pro detekci instalovaných modulů a programů, po splnění jejich požadavků programy fungují správně.

V Praze, 25. 8. 2011
RNDr. Miroslav Spousta, ÚFAL